Visceral versus Verbal: Can We See Depression?

Xuanying Zhu, Tom Gedeon, Sabrina Caldwell, Richard Jones

Research School of Computer Science, The Australian National University, Canberra, Australia 2601

E-mail: xuanying.zhu@anu.edu.au, tom@cs.anu.edu.au, sabrina.caldwell@anu.edu.au, richard.jones@anu.edu.au

Abstract: Depression widely affects global populations and is one of the leading causes of disability and suicide. Despite its prevalence, traditional diagnosis for depression is exceedingly associated with misidentification and over-estimation, due to its subjective nature. With advances in affective computing, computational approaches make it possible to discern depression through second party physiological indicators; people observing the behaviour of depressed individuals have measurable changes in their physiological signals. We explored Blood volume pulse (BVP), Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD) from observers as valid sources to indicate depression in others. The behaviour of individuals suffering from four levels of depression was shown in 16 videos to 12 experimental observers whose physiological signals were recorded. We found that depression provokes visceral physiological reactions in observers that we can measure, resulting in neural network classification of 94% accuracy. In contrast, we also found that depression does not provoke strong conscious recognition ('verbal') in observers, which is only slightly over a chance level, at 27%.

Keywords: depression detection; physiological signals; observers; galvanic skin response; skin temperature; blood volume pulse; pupillary dilation, affective computing

1 Introduction

Major depressive disorder, or 'depression' for short, is a common but serious mental disorder that widely affects populations around the world [1]. Its cause is believed to be a combination of genetics [2] and environmental factors [3], such as, major life changes, trauma, or long-lasting exposure to difficulties. It usually presents with persistent depressed mood, loss of interest and enjoyment, feelings of sadness, guilt or low self-esteem, poor concentration, and at its worst, suicidal actions [4]. According to the World Health Organization (WHO) [1], depression is one of the leading causes of disability, affecting more than 300 million people.

Since depression comes with some observable behavioural symptoms regarding the normal expression of emotions and general functioning [5], traditional diagnostic approaches for depression rely on subjective measures of behaviours. These methods are typically involved with self-reported questionnaires such as the Beck Depression Index (BDI) [6], or clinician-assisted interview style assessments such as the Hamilton Rating Scale for Depression (HAMD) [7], which score patients' depression level by the severity of their symptoms. However, meta-analyses of depression diagnosis have indicated the wide-spread existence of both over- and under-recognition [8] [9]. The central problem is that these diagnostic tools are subjective and biased, as they are heavily associated with patients' sensitivity to symptoms and willingness to honestly reveal the symptoms [10]. Given that the accuracy of depression diagnosis correlates with reassessments and longer consultation time [8], these approaches can be time-consuming. To better serve the needs of the patient, medical profession and community, it is desirable that we find simpler and less subjective methods of depression diagnosis.

To tackle the unreliable issues of subjective assessments of depression and other emotions, research has explored the possibility of measuring emotions objectively via human physiological signals along with self-assessment reports [11], based on the demonstration that physiological signals are highly correlated with subject assessments [12]. These measures are typically automated and involve affective sensors to study changes in galvanic skin response (GSR) [13], blood volume and heart rate activities [13]–[16], pupillary sizes and eye movements [17].

That emotion can be distinguished based on physiological signals relies on the human peripheral nervous system, which consists of, the somatic nervous system (SoNS), which controls voluntary body movement, and the autonomic nervous system (ANS). The ANS is responsible for involuntary activities and, without conscious awareness, it automatically regulates bodily functions such as heart rate, respiratory rate, and pupillary responses. The ANS consists of the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The PNS is responsible for activities during resting and digesting states. The SNS, dominates when a person is threatened or under stress. To get the body ready to cope with danger, it expands pupils, allowing more light to enter the eyes for better vision, and increases the respiratory rate and heart rate, providing better oxygenation and easier blood flow throughout the body. It is often known as the 'fight or flight' system. Thus, we expect that SNS activation will change multiple physiological signals in the individuals undergoing such emotion.

Given that physiological responses maintained by the ANS can indicate individuals' inner states without conscious awareness, physiological signals have been used as biomarkers for depression. For example, depressed patients have different eye gaze behaviours [18], lower Galvanic Skin Response (GSR) [19], reduced Heart Rate Variability (HRV) [20], and depressed brain activities as captured by Near Infrared Spectroscopy (NIRS) [21] and Electroencephalogram (EEG) [22]. These signals provide more objective and quantitative criteria, and

when combined with machine learning technologies, can play an essential role in providing an objective assessment for depression.

Despite the physiological correlates of depressed patients, since individuals with depression tend to withdraw from social activities [1] [8], and so leaving them with less chance to access these facilities, our goal is to investigate physiological signals of *observers* to identify others' depression level. Our previous work demonstrated the feasibility of using observers' physiological signals as indicators of other individuals' depression [23] using neural networks. Subtle cues could be noticed by observers, which are reflected in observers' physiological signals. We found that neural networks trained with physiological features can recognise other individuals' depression levels with 92% accuracy. We extend this methodological and analytical approach, to ascertain whether neural networks trained on observers' BVP and associated heart rate (HR) and heart rate variability (HRV) signals can identify other individuals' depression. The identification of universal physiological indicators from observers watching depressed individuals could assist with earlier diagnosis, which, combined with known effective treatments, would decrease the burden for individuals and society. The use of physiological signals could also be applicable in other domains such as engineering [24].

This paper examines whether observers' BVP and associated heart rate (HR) and heart rate variability (HRV) signals respond to depressed individuals and whether a computational model trained with single BVP signal, as well as, trained with a hybrid of four physiological signals (GSR, BVP, ST, and PD), could better recognise other individuals' depression level. It details an experiment conducted to collect multiple physiological response signals from experiment participants who watched videos of people with various levels of depression and includes selecting optimally useful features from the response signals. The paper concludes with a summary of the findings and suggests directions for future work.

2 Experimental Design

Our aim is to detect other individuals' depression using observers' Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Skin Temperature (ST) and Pupillary Dilation (PD) signals, both singly and in combination. Following a similar experimental design to our previous work [23], we selected sixteen videos from the 2014 Audio-Visual Emotion Challenge (AVEC 2014) dataset as stimuli [25] in which individuals with four depression severities read aloud a paragraph in German. We recruited 12 participants as observers to watch the video stimuli, while we recorded their BVP, GSR, ST, and PD. We also collected observers' conscious subjective depression prediction of the individuals in the videos via a survey. A schematic diagram of the equipment setup is provided in Figure 1.

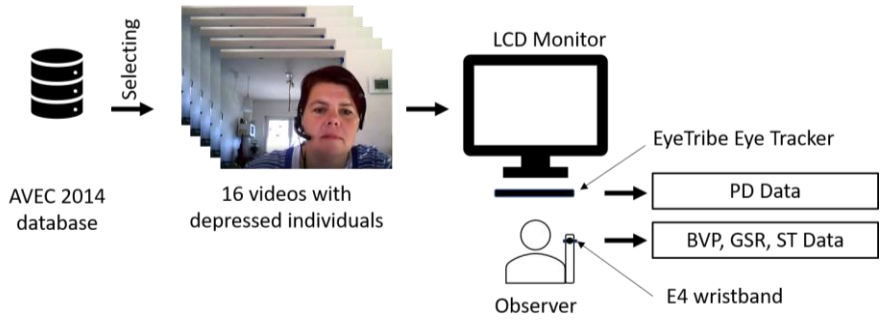

Figure 1
A schematic diagram of the equipment setup

2.1 Stimuli

We used videos from the Northwind category of the AVEC 2014 dataset [25]. The Northwind category consists of 150 webcam video recordings of 4 categories of participants individually reading aloud a paragraph in German. Each recording was labelled with a single depression level derived from involved participants self-reported depression level indicated by the Beck Depression Inventory – II (BDI-II) [6]. This index gives depression scores ranging from 0 to 63 and groups the scores into four depression categories:

- 0 - 13 Indicates no or minimal depression
- 14 - 19 Mild depression
- 20 - 28 Moderate depression
- 29 - 63 Severe depression

We chose 16 videos (see Table 1) with similar durations, from 36 s to 50 s (Ave = 41.2, Standard Deviation = 3.8), evenly across the four depression categories.

Table 1
Stimuli videos selected from the testing set of Northwind tasks in AVEC 2014

Video name	Duration (sec)	Depression level	Category
210_2, 249_1, 341_1, 240_3	43, 42, 39, 41	1, 4, 7, 11	no
220_3, 242_1, 315_3, 214_3	39, 42, 40 43	15, 16, 17, 18	mild
245_3, 218_3, 325_2. 250_1	40, 39, 39, 41	21	moderate
226_2, 359_1. 315_2, 237_1	41, 45, 58, 47	30	severe

2.2 Participants

Fourteen students who do not understand German and do not have prior training in depression recognition took part in the experiment. They were recruited to watch German-language depression videos as §2.4Procedure below describes. Ethics Approval was obtained from the Australian National University Human Research Ethics Committee. Two subjects were excluded based on the predefined exclusion criteria for having a history of cardiovascular disease or technical failures of the sensors. The final sample consisted of 12 participants, six males, and six females, from 18 to 27 years in age (mean = 21.1, standard deviation = 2.8) with normal or corrected-to-normal vision and hearing. This sample size of participants is normal for publications as a preliminary study in medicine [26].

2.3 Measures and Sensors

2.3.1 Blood Volume Pulse (BVP)

Blood Volume Pulse (BVP) indicates the volume of blood running through the vessels over time [27]. It can be measured by a photoplethysmographic (PPG) sensor using infra-red light through the skin surface and measures the reflected light. With every beat, the heart pushes a volume of blood causing a wave which travels from the heart, and returns to the heart. As the surge of blood dissipates, the signal falls. The direct pulse wave then bounces back from the lower body, causing a secondary wave, which appears as a second rise in the signal. The signal then drops until the next heartbeat. A typical BVP signal is illustrated in Figure 2, which consists of the systolic peak (Figure 2-I), dicrotic notch (2-II), diastolic peak (2-III) and diastolic point (2-IV).

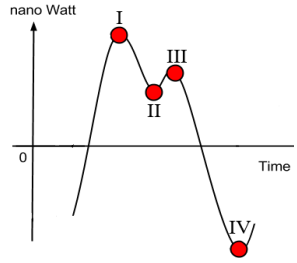

Figure 2
A typical BVP waveform

BVP can provide information on changes in SNS activation, which are influenced by emotional context. For example, BVP is negatively correlated with stress and positively correlated with sadness [28]. We can derive other cardiovascular measures from BVP, such as heart rate (HR) and heart rate variability (HRV). HR is a useful predictor of emotional valence and can distinguish between positive and negative emotions [29]. HRV refers to the temporal, beat-to-beat variations, in the consecutive heartbeats, and can indicate mental effort and emotions [28].

We placed an Empatica E4 wristband on the wrist of the non-dominant hand of observers [30], which recorded BVP with a sampling rate of 64 Hz [31].

2.3.2 Galvanic Skin Response (GSR)

Galvanic Skin Response (GSR) measures electrical conductivity of the skin, which varies due to the amount of sweat [32]. Stress or danger stimulates glands to produce salty sweat, which increases skin conductivity [33]. GSR is composed of two separate electro dermal activities: the tonic component is slow-moving and shows the general activity of the perspiratory glands caused by body or external temperature, while the phasic component is a faster distinctive waveform in the signal, and is considered to be linearly correlated with the intensity of arousal in mental state [11]. In this study, we recorded participants' wrist GSR using an Empatica E4 wristband with a sampling rate of 4 Hz [31].

2.3.3 Skin Temperature (ST)

Skin Temperature (ST) fluctuates due to vasodilatation of peripheral blood vessels induced by increased activity of the SNS. It is negatively correlated with unpleasant emotions such as stress [34] and fear [35] because blood is redirected to vital organs as a protection measure. In this study, we recorded participants' wrist ST using an Empatica E4 wristband with a sampling rate of 4 Hz [31].

2.3.4 Pupillary Dilation (PD)

Pupillary Dilation (PD) provides indications of changes in mental states and of mental activities [36]. Pupil size was found to respond to emotionally stimuli. The pupil is significantly bigger after positively or negatively arousing stimuli than after neutral stimuli [37]. We used The EyeTribe, an affordable, non-intrusive and precise eye tracker [38], to record pupil size at 60 Hz. Python code was written to analyse data collected by the EyeTribe SDK software [39].

2.4 Procedure

The experiment was conducted with each participant in the same quiet experiment room. Participants were given a written set of instructions and guidance from the experiment instructor before they provided written informed consent. An Empatica E4 sensor [31] was attached to the wrist of each participant's non-dominant hand [30], and eye gaze calibration for the eye tracker was performed.

Participants then filled in a questionnaire to collect demographic and health characteristics that may affect cardiovascular and pupillary responses. Each participant then watched 16 videos and was asked at the end of each video to respond to a question of "How would you like to rank the patient's depression level?" on a four-item scale of "None, Mild, Moderate, Severe" that matches with the BDI-II [6] scale. A five-second gap was provided between videos. The videos were presented in an order balanced way to avoid the effects of presentation order. At the end of the experiment, participants filled in the BDI-II [6] survey assessing their depression level. In total, the experiment took approximately forty minutes.

3 Methodology

Following our previous work methodology [23], we first pre-processed the physiological responses of observers to remove noise and individual bias. We then computed features for the four recorded physiological signals before we trained neural networks with the most significant features selected by a genetic algorithm. An overall structure of our depression recognition system is illustrated in Figure 3.

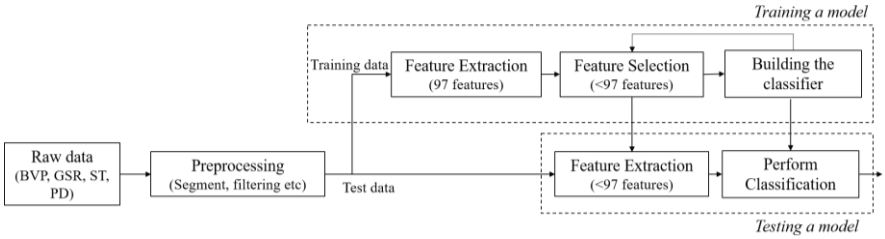

Figure 3
An overall structure of our depression recognition system

3.1 Pre-processing

For all four physiological signals, we first extracted the raw signal for all observers when they were watching the full set of 16 videos, removing the noise caused by the movement of observers, which mostly happened at the beginning and the end of the recording when they were filling in the demographic questionnaire and post-experiment survey. We then applied a cubic spline interpolation to the missing pupil size data caused by occasional eye blinks. This procedure was employed on the left and the right pupil data separately.

To reduce the between-participant differences, we then separately normalised BVP, GSR, ST, left and right PD to the range between 0 and 1 with a min-max normalisation scaler as shown in (1):

$$S_{normalised} = \frac{S - \min(S)}{\max(S) - \min(S)} \tag{1}$$

Where $S_{normalised}$ is the min-max scaled data of signal S, and $\max(S)$ and $\min(S)$ are the maximum and minimum value of signal S.

After normalisation, to remove noise artefacts, we applied a lowpass Butterworth filter to BVP, GSR and ST with an order of 6 and a cut-off frequency of 0.5 Hz, 0.2 Hz [13] and 0.3 Hz [40] to form the Low Pass (LP) BVP, LP GSR, and LP ST data, respectively. We also filtered the left and right PD data with a 10-point Hann moving window. The average pupillary size of the normalised left and right pupil data was then calculated.

Following this, we further segmented both the normalised and filtered signals by each video watching session, so that each segmented physiological data set corresponds to one observer's physiological state invoked by his or her experience of watching one video.

3.2 Feature Extraction

After pre-processing the raw signals, we generated time- and frequency-domain features that characterise the changes in the physiological signals over the time observer participants spent on watching each video.

3.2.1 Blood Volume Pulse (BVP) Features

According to the literature of using BVP for emotion recognition [41], we first calculated the following six time-domain features from the LP BVP.

1) Minimum
2) Maximum
3) Mean
4) Standard deviation
5) Variance
6) Root mean square

Let R_i be the i^{th} systolic peak, RR_i be the interval between peak R_{i+1} and R_i, and RR_{diff} i be the differences between intervals RR_{i+1} and RR_i (as Figure 4 shows). Heartbeats defined as the systolic peaks of the LP BVP as illustrated as R_n in Figure 4 were then identified adapting a peak detection technique devised by Van Gent et al [42]. We calculated a moving average using a window of 0.8 seconds before and after each data point. Regions of Interest (ROI) are then marked between two diastolic points where the amplitude of the signal is larger than the moving average. Systolic peaks were detected at the maximum of each ROI.

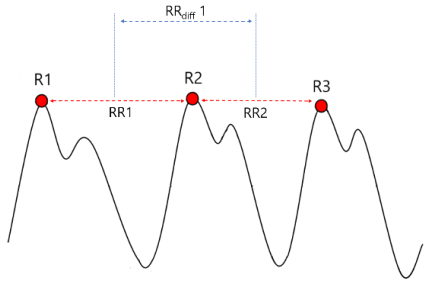

Figure 4
BVP signal with systolic peaks, peak intervals and differences between intervals annotated

To extract time-domain heart rate features, or heart rate variability, we computed the following 8 time-domain features that were previously shown to be correlated with external stimuli and mental states [43]–[45].

Inter beats interval (IBI) $$IBI = \overline{RR} \quad (2)$$

Ave. beats per min (BPM) $\text{BPM} = (60 \times \text{Sampling Rate}) / IBI$ (3)

Standard deviation of intervals between heart beats (SDNN)

$$SDNN = \sqrt{\frac{\sum_{n=1}^{N}\left(RR_i - \overline{RR}\right)}{N-1}} \quad (4)$$

Standard deviation of differences of adjacent R-R intervals (SDSD)

$$SDSD = \sqrt{\frac{\sum_{n=1}^{N}\left(RR_{diff_i} - \overline{RR_{diff}}\right)^2}{N-1}} \quad (4)$$

Root mean square of differences of adjacent R-R intervals (rMSSD)

$$rMMSSD = \sqrt{\frac{\sum_{n=1}^{N}\left(RR_{diff_i}\right)^2}{N-1}} \quad (5)$$

Percentage of the differences greater than 20 ms (pNN20)

Percentage of the differences greater than 50 ms (pNN50)

Proportion of differences greater than 50 ms / 20 ms (pNN50/pNN20)

As indicated in [46], the spectral features of heart rate signal are more robust for short time durations and less sensitive to missing heartbeats, so we also included 5 frequency-domain features after we performed a Fast Fourier Transform (FFT) over the peak intervals to convert the signal into the frequency domain.

High frequency power (HFP): ranging from 0.15 to 0.5 Hz

Low frequency power (LFP): in a range between 0.04 and 0.15 Hz

Very low frequency power (VLFP) in a range between 0.003 and 0.04 Hz

LF/HF ratio (LFHF): the ratio of LFP over HFP

Respiratory Rate (RSP): max power in frequency range of 0.1-0.25 Hz

3.2.2 Galvanic Skin Response (GSR) Features

Sixteen time-domain features were calculated from both the normalised GSR and LP GSR separately, based on the following statistical methods:

a) Minimum
b) Maximum
c) Mean
d) Standard deviation
e) Variance
f) Root mean square
g) Mean of absolute values of first difference
h) Mean of absolute values of second difference

GSR consists of a tonic component (also called DC level) and a phasic component (also called the skin conductance response, SCR) [11]. DC level shows the long-term slow variation in the signal, indicating general activity of perspiratory glands caused by body or external temperature, while SCR reflects relatively faster responses to external stimuli. To extract the DC level component, we used a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to obtain the Very Low Pass signal (VLP). We further acquired a detrended SCR signal without DC component by removing continuous piecewise linear trend in both LP and VLP signal. Afterward, we calculated the following frequency-domain features:

- Number of SCR occurrences for VLP, LP and normalised GSR
- Mean of amplitudes of SCRs for VLP, LP and normalised GSR
- Ratio of SCR occurrences in VLP to occurrences in LP

3.2.3 Skin Temperature (ST) Features

For ST, we used a similar feature extraction approach as for the GSR signal. We calculated 16 time-domain features which include the minimum, maximum, mean, standard deviation, variance, root mean square, means of the absolute values of the first and second difference of the normalised and LP ST signal. Subsequently, we applied a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to the normalised ST signal to form the VLP ST signal. We finally calculated the numbers and amplitudes of peak occurrences for VLP and LP ST signals as well as the ratio of peak occurrences in VLP to those in LP as features.

3.2.4 Pupillary Dilation (PD) Features

Similar to the GSR and ST signals, for PD, we first calculated the following 8 features from the normalised left PD, normalised right PD, and average PD separately: minimum, maximum, mean, standard deviation, variance, root mean square, means of the absolute values of the first difference and means of the absolute values of the second difference. We then applied a very low pass Butterworth filter with a cut-off frequency of 0.08 Hz to the normalised left, right and average PD signal to form the left VLP PD, right VLP PD, and average VLP PD. Numbers and amplitudes of peak occurrences for left, right and average VLP and LP PD signals, as well as the ratio of peak occurrences in VLP to those in LP for the left, right and average signals, were subsequently extracted as features. Thus, we collected a total of 104 features from the four physiological signals: 19 (BVP) + 23 (GSR) + 23 (ST) + 39 (PD).

3.3 Feature Selection

Our previous work [23] indicated that neural networks (NNs) trained with subsets of features selected by Genetic Algorithm (GA) [47] perform better at depression prediction than without feature selection. A full set of features may include

redundant / irrelevant features that outweigh more useful features. To make a direct comparison to our prior work, we used a GA feature selection method.

The initial population for the GA was set to use all features. A candidate chromosome was defined as a binary string where the index for a bit represented a feature, and the bit value indicated whether the feature was used for classification. The presence (1) or absence (0) of every possible feature was determined based on a fitness function, which is the depression recognition performance of an NN. An example of such representation is demonstrated in Figure 5. All settings for the GA used in the hybrid classification system can be found in Table 2.

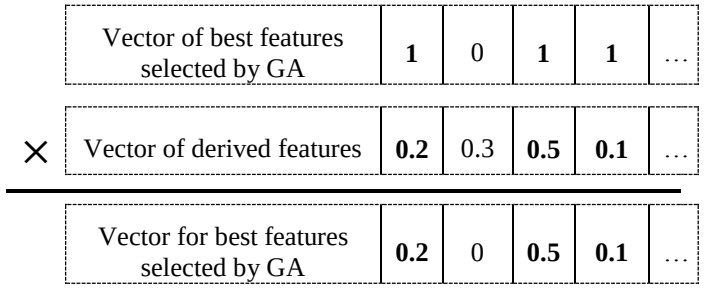

Figure 5
GA representation of features

Table 2
GA settings

GA Parameter	Value
Population size	100
Crossover rate	0.8
Mutation rate	1/(length of the chromosome)
Crossover type	Uniform crossover
Mutation type	Uniform mutation
Selection type	Stochastic uniform selection

3.4 Neural Network Classifiers

In this study, we were interested in determining the depression recognition capability achieved by a combination of BVP, GSR, ST, and PD measurements as monitored signals. Assessment of overall usefulness of signals is also important as fewer sensors are required if only a single signal is needed to achieve similar recognition capability. Thus, we trained five NN classifiers with the following five conditions: 1) BVP+GSR+ST+PD: using a subset of features selected by GA from all features extracted from all four signals; 2-5) each of BVP/GSR/ST/PD singly: using a subset of features selected by GA from features extracted from the

BVP/GSR/ST/PD signal. We note that when each physiological signal was used, the classifier was retrained and retested using the same validation scheme.

All five NNs performed a 4-class classification indicating 4 depression severities using 4 output neurons. The first NN was set to have a sigmoid hidden layer of 100 neurons after we tested the first NN with different hidden neuron size from 10 to 200 and found 100 to be optimal. With a similar approach, the other four NNs use 50 neurons. All NNs were trained with a commonly used optimizer, the Adam optimizer [48] using backpropagation with the Cross-Entropy loss function.

As noted in [23], that in the context of continuous physiological data, training a classifier on random splits of data is not appropriate, we used the leave-one-participant-out validation method. For each run, we took physiological features from one observer as the testing set, and those from the remaining participants as the training set. We repeated this process for all observers, each time leaving out physiological features from a different observer as the testing set. We averaged the performance as the final results reported.

3.5 Evaluation Measures

To validate the effectiveness of our models, we used *precision*, *recall*, and *F1-score* as evaluation measures. For a specific depression level L, *precision* is defined as the proportion of individuals that are correctly predicted with depression level L and actually have that level of depression; *recall* is the percentage of depressed individuals that are correctly predicted with depression level L among all individuals labelled with depression level L; and *F1-score* takes the harmonic mean of precision and recall defined as $2 \times \frac{\text{Precision}_\text{L} \times Recall_L}{Precision_L + Recall_L}$.

As multiclass depression labels were also predicted by our models, we calculated the average precision, recall, and F1 score for all depression levels as a whole, to give a view on the general prediction performance. Also, we computed the overall accuracy, which is the number of individuals correctly predicted with their corresponding depression levels by the model over total number of individuals.

4 Results and Discussion

4.1 Observers Subjective Prediction

As Table 3 shows, observers are not good at *consciously* identifying the depression severity of other individuals in videos. The overall accuracy was 27%, which is slightly over the *prima-facie* chance level of 25% since there were four options for observers over balanced numbers of video stimuli. This is consistent

with earlier findings about the accuracy of people's conscious judgments on the veracity of smiles [49], anger [50] and deceiving behaviours [51], which are all only marginally higher than chance level. Thus a prediction barely over chance in general is not surprising. However, the low recognition accuracy could be exacerbated by our observer participants being recruited from a naïve population who had not previously received any training regarding depression diagnosis. Future research should explore the accuracy of conscious judgments from psychologists who are trained to diagnose depression patients.

Table 3
Results of depression prediction from observers' subjective verbal responses

Depression level	**Subjective Prediction**		
	Precision	***Recall***	***F1 score***
None	0.31	0.33	0.32
Mild	0.18	0.21	0.19
Moderate	0.23	0.25	0.24
Severe	0.42	0.29	0.35
Average	0.29	0.27	0.28
Overall Accuracy	0.27		

The average ratios of consciously identifying healthy individuals and severely depressed individuals correctly were 33% and 29% respectively, higher than those of identifying depressed individuals in the middle ranges, at 21% and 25%. This could imply that people are better at identifying healthy individuals and depressed patients with severe symptoms, but worse at differentiating depression levels.

4.2 Classification based on All Physiological Signals

All features derived from observers' BVP, GSR, ST, and PD signals were provided to NNs with GA for feature selection. Performance of the classifications was calculated based on the average results of 10 runs and shown in Table 4.

Table 4
Results of depression prediction from NN trained with BVP, GSR, ST, and PD features selected by GA compared with our previous study trained with GSR, ST and PD features selected by GA

Depression level	**Our previous study** [23]			**This study**		
	Precision	***Recall***	***F1 score***	***Precision***	***Recall***	***F1 score***
None	0.92	0.95	0.94	0.90	0.98	0.94
Mild	0.93	0.89	0.91	0.94	0.92	0.93
Moderate	0.88	0.90	0.89	0.95	0.92	0.93
Severe	0.95	0.95	0.95	0.98	0.96	0.97
Average	0.92	0.92	0.92	0.94	0.94	0.94
Overall Accuracy	0.92			0.94		

As can be seen above, the overall prediction accuracy across all four depression levels was much higher than observers' conscious judgments, at 94%. Statistical analysis was conducted on the results using the Student's t-test since models trained with different physiological features share normality and equality of variances across comparison groups. In accordance with the Student's t-test, the model trained with BVP, GSR, ST, and PD features produced significantly better depression recognition rates than the conscious evaluation of observers ($p < 0.005$). This could indicate that although humans are not good at consciously detecting the depression severity of others, they can emotionally sense depression in others. Their physiological changes provoked by depression from others can be effectively detected by computational classifiers such as neural networks. The superior detecting ability of human unconscious physiological responses over conscious judgments is also found in other research in which the realness of two basic emotions are examined [22] [23]. Taken together, it could suggest that unconscious responses from instinctive human ability, which has been adaptively evolved by natural selection, can make effective use of cues to identify depressed individuals without being influenced by conscious biases.

Compared to our previous work, where only three physiological signals, GSR, ST, and PD were examined, our improved model received a statistically significantly better overall accuracy ($p < 0.01$). Our improved model also outperformed in recognising depression at mild, moderate, and severe levels with higher F1 scores ($p < 0.01$). This may imply that BVP and associated HR and HRV signals improve identifying depression severity of depressed individuals when combined with other physiological signals.

However, this model obtained a slightly less precision rate when the video individuals do not have depression, meaning that fewer video individuals with no depression were correctly identified with no depression and thus more video individuals with no depression were overestimated to have some levels of depression. It could indicate that while BVP improves the recognition rate of depressed individuals, it may also over-recognise depression in healthy individuals. This should be investigated and overcome for clinical applications.

Further, similar to clinical depression diagnosis where the middle levels of depression are harder to correctly identify [52], our model performed slightly worse in predicting depression levels of individuals with mild and moderate depression severity than those with none and severe depression level, reflected by the lower F1-scores of mild and moderate depression level. This result is also found in observers' subjective predictions and our previous work [23] [53], indicating the difficulty of accurately recognising middle levels of depression. Future research can consider exploring the feasibility of analysing more complex physiological signals, such as brain activity tracking with electroencephalogram (EEG) [54]–[56] or functional near-infrared spectroscopy (fNIRS) [52], to detect more subtle cues in observed stimuli.

4.3 Classification based on Individual Physiological Signal

To evaluate the classification capability of models with fewer physiological signals, features derived from individual physiological signals were provided to the NN model with GA feature selection. Performance of the classifications were calculated based on the average results of 10 runs and presented in Table 5.

Table 5
Results of depression prediction from NN trained with single physiological signal with GA

Depression level	**BVP**			**GSR**			**ST**			**PD**		
	Precision	***Recall***	***F1 score***	***Precision***	Recall	F1 score	***Precision***	Recall	F1 score	***Precision***	***Recall***	***F1 score***
None	0.90	0.81	0.85	**0.92**	**0.94**	**0.93**	0.87	0.83	0.85	0.91	0.92	0.91
Mild	0.84	**0.92**	0.88	0.89	0.81	0.85	0.84	0.83	0.84	**0.92**	0.91	**0.91**
Moderate	0.90	**0.95**	0.92	0.87	0.87	0.87	0.80	0.81	0.81	**0.93**	0.92	**0.92**
Severe	0.90	0.86	0.88	0.87	**0.93**	0.90	0.84	0.86	0.85	**0.95**	0.92	**0.94**
Average	0.89	0.89	0.88	0.89	0.89	0.89	0.84	0.83	0.84	**0.93**	**0.92**	**0.93**
Overall Accuracy	0.89			0.89			0.84			**0.93**		

When features from only one physiological signal were available, depression patterns were best recognised from PD features, with an average F1 score and an overall accuracy of 93%. These results were ≥4% higher than models trained with other individual signals. BVP features and GSR features contributed similarly to depression recognition in general, while ST features were less accurate. We also performed statistical analyses on the average F1 score and the overall accuracy of each pair of models trained with different single physiological signals separately. The Student's t-test has shown a significant difference between models trained with PD and models trained with BVP, GSR, or ST ($p < 0.01$). This is consistent with the literature [57] where pupil size was prominent among other signals in detecting stress, revealing some physiological signals convey more informative features to classifiers. It also indicated that models trained with ST features were significantly less likely to predict depression level ($p < 0.01$). No significance was found between models trained with BVP and GSR features ($p > 0.1$).

The contribution of each signal to the prediction of each depression level can also be seen in Table **5** based on the precision, recall, and F1 score of each depression level. PD was the best in recognising all depression levels except the "None" category, achieved by having the highest precision and F1 score ($p<0.005$). On the other hand, when identifying individuals with no depression levels, GSR obtained the best result across precision, recall, and F1 score ($p<0.05$). Taken together, it possibly shows that while PD is a valid indicator of other individuals' depression state, GSR is more useful to recognise healthy individuals.

Although models trained with PD-only features obtained an acceptable recognition performance, when compared with the models trained with a hybrid of four physiological signals, BVP, GSR, ST and PD, it performed slightly worse,

reflected by lower precision rates, recall rates and F1 scores across all four depression levels. Statistical significance was found for recognition performances of mild, moderate, and severe depression between PD and a hybrid of 4 signals ($p < 0.05$) but no significance was found when observed individuals were healthy. This phenomenon also happened in [58] [59] where a combination of multiple signals outperforms models trained with individual signals separately. It is probably because we as human beings use a combination of different modalities in our body to express emotions and thus our physiological signals have been evolving and favoured by natural selection to function as a whole [60].

6 Limitations, Future Work and Conclusions

Observers in this study are naïve individuals who do not have depression diagnosis experience and do not understand German, which is the language spoken by the individual in the videos. Future research should recruit clinicians who are skilled in diagnosing depression and German speakers to evaluate the effect of domain knowledge and language understanding on depression prediction. Stronger conclusions could be drawn in subsequent studies, with more observers involved. Different neural network settings and structures could also be explored. Other directions for future work include studying how generalizable our results are in more realistic environments such as in daily social interactions. Finally, use of more complex physiological signals, such as, brain activity tracking, with EEG and fNIRS, could also be investigated, to perform better recognition.

Conclusions

In this article, we explored the use of physiological signals from observers, to detect depression severity, of other individuals. We investigated the utility of BVP, GSR, ST, and PD from observers (singly and in combination) to predict the depression level of individuals they observed in videos. The results show that the combination of these four signals achieved an NN classification accuracy of 94%, outperforming models trained in our previous work, which did not include BVP and associated HR and HRV signals and models trained using individual signals separately. We also identified PD as the most promising physiological source for depression recognition, as well as, the potential of GSR for recognising healthy individuals among depressed patients. Future research and implementation of the findings in this area are likely to be beneficial in assisting with more objective and earlier depression diagnosis, which combined with the use of known effective treatments, could decrease the burden of depression for individuals and society.

References

[1] M. Marcus, M. T. Yasamy, M. van Ommeren, D. Chisholm, and S. Saxena, "Depression: A Global Public Health Concern," *WHO Dep. Ment. Heal. Subst. Abus.*, 2012

[2] R. Chandra *et al.*, "Reduced Slc6a15 in nucleus accumbens D2-neurons underlies stress susceptibility," *J. Neurosci.*, Vol. 37, No. 27, pp. 6527-6538, 2017

[3] A. T. Beck and B. A. Alford, *Depression: Causes and treatment*. University of Pennsylvania Press, 2009

[4] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," *J. Affect. Disord.*, Vol. 147, No. 1-3, pp. 17-28, 2013

[5] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Curr. Opin. Psychol.*, Vol. 4, pp. 75-79, 2015

[6] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," *San Antonio*, Vol. 78, No. 2, pp. 490-498, 1996

[7] M. Hamilton, "A rating scale for depression," *J. Neurol. Neurosurg. Psychiatry*, Vol. 23, No. 1, p. 56, 1960

[8] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *Lancet*, Vol. 374, No. 9690, pp. 609-619, 2009

[9] J. A. Bilello, "Seeking an objective diagnosis of depression," *Biomark. Med.*, Vol. 10, No. 8, pp. 861-875, 2016

[10] J. Joshi *et al.*, "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Interfaces*, Vol. 7, No. 3, pp. 217-228, 2013

[11] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 30, No. 12, pp. 2067-2083, 2008

[12] J. Tomaka, J. Blascovich, R. M. Kelsey, and C. L. Leitten, "Subjective, physiological, and behavioral effects of threat and challenge appraisal.," *J. Pers. Soc. Psychol.*, Vol. 65, No. 2, p. 248, 1993

[13] K. Hercegfi, "Heart rate variability monitoring during human-computer interaction," *Acta Polytech. Hungarica*, Vol. 8, No. 5, pp. 205-224, 2011

[14] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Tutorial and research workshop on affective dialogue systems*, 2004, pp. 36-48

[15] S. Khoór *et al.*, "Heart Rate Analysis and Telemedicine: New Concepts & Maths," *Acta Polytech. Hungarica*, Vol. 5, No. 1, 2008

[16] G. Förd\Hos, I. Bosznai, L. Kovács, B. Benyó, and Z. Benyó, "Sensor-net for monitoring vital parameters of vehicle drivers," *ACTA Polytech. hungarica*, Vol. 4, No. 4, pp. 25-36, 2007

[17] P. C. Schmid, M. S. Mast, D. Bombari, F. W. Mast, and J. S. Lobmaier, "How mood states affect information processing during facial emotion recognition: an eye tracking study," *Swiss J. Psychol.*, 2011

[18] S. Scherer *et al.*, "Automatic behavior descriptors for psychological disorder analysis," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013, pp. 1-8

[19] Y.-T. Chen, I.-C. Hung, M.-W. Huang, C.-J. Hou, and K.-S. Cheng, "Physiological signal analysis for patients with depression," in *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, 2011, Vol. 2, pp. 805-808

[20] F. A. Jain *et al.*, "Heart rate variability and treatment outcome in major depression: a pilot study," *Int. J. Psychophysiol.*, Vol. 93, No. 2, pp. 204-210, 2014

[21] T. Fekete, F. D. C. C. Beacher, J. Cha, D. Rubin, and L. R. Mujica-Parodi, "Small-world network properties in prefrontal cortex correlate with predictors of psychopathology risk in young children: A NIRS study," *Neuroimage*, Vol. 85, pp. 345-353, 2014

[22] N. V Thakor and S. Tong, "Advances in quantitative electroencephalogram analysis methods," *Annu. Rev. Biomed. Eng.*, Vol. 6, pp. 453-495, 2004

[23] X. Zhu, T. Gedeon, S. Caldwell, and R. Jones, "Detecting emotional reactions to videos of depression," in *IEEE International Conference on Intelligent Engineering Systems*, 2019

[24] J. Katona and A. Kovari, "A Brain--Computer Interface Project Applied in Computer Engineering," *IEEE Trans. Educ.*, Vol. 59, No. 4, pp. 319-326, 2016

[25] M. Valstar *et al.*, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," *Proc. 4th ACM Int. Work. Audio/Visual Emot. Chall. (AVEC '14)* pp. 3-10, 2014

[26] R. Simon, "Optimal two-stage designs for phase II clinical trials," *Contemp. Clin. Trials*, Vol. 10, No. 1, pp. 1-10, 1989

[27] K. Gouizi, F. Bereksi Reguig, and C. Maaoui, "Emotion recognition from physiological signals," *J. Med. Eng. Technol.*, Vol. 35, No. 7, pp. 300-307, 2011

[28] N. Sharma and T. Gedeon, "Modeling a stress signal," *Appl. Soft Comput.*, Vol. 14, pp. 53-61, 2014

[29] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, Vol. 6, No. March 2014, pp. 37-41, 2008

[30] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," *Handb. Psychophysiol.*, Vol. 2, pp. 200-223, 2007

[31] Empatica, "E4 wristband." [Online] Available: https://www.empatica.com/research/e4/ [Accessed: 30-May-2018]

[32] R. M. Stern, W. J. Ray, and K. S. Quigley, *Psychophysiological recording*. Oxford University Press, USA, 2001

[33] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012

[34] F. Al-Shargie, T. B. Tang, and M. Kiguchi, "Mental stress grading based on fNIRS signals," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2016, Vol. 2016-Octob, pp. 5140-5143

[35] J. E. LeDoux and S. G. Hofmann, "The subjective experience of emotion: a fearful view," *Curr. Opin. Behav. Sci.*, Vol. 19, pp. 67-72, Feb. 2018

[36] M. Porta, S. Ricotti, and C. J. Perez, "Emotional e-learning through eye tracking," in *Global Engineering Education Conference (EDUCON) 2012 IEEE*, 2012, pp. 1-6

[37] B. Laeng, S. Sirois, and G. Gredebäck, "Pupillometry: a window to the preconscious?," *Perspect. Psychol. Sci.*, Vol. 7, No. 1, pp. 18-27, 2012

[38] E. Dalmaijer, "Is the low-cost EyeTribe eye tracker any good for research?," 2014

[39] TheEyeTribe, "The EyeTribe." [Online] Available: http://theeyetribe.com/theeyetribe.com/about/index.html

[40] V. Xia, N. Jaques, S. Taylor, S. Fedor, and R. Picard, "Active learning for electrodermal activity classification," in *Signal Processing in Medicine and Biology Symposium (SPMB) 2015 IEEE*, 2015, pp. 1-6

[41] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau, "Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites," *Physiol. Meas.*, Vol. 32, No. 10, p. 1529, 2011

[42] P. Van Gent, H. Farah, N. van Nes, and B. van Arem, "Analysing Noisy Driver Physiology Real-Time Using Off-the-Self Sensors: Heart Rate Analysis Software from the Taking the Fast Lane Project.," *J. Open Res. Softw.*, 2018

[43] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Rev. Gen. Psychol.*, Vol. 10, No. 3, pp. 229-240, 2006

[44] M. Züger and T. Fritz, "Interruptibility of software developers and its prediction using psycho-physiological sensors," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2981-2990

[45] C. G. Beevers, A. J. Ellis, and R. M. Reid, "Heart rate variability predicts cognitive reactivity to a sad mood provocation," *Cognit. Ther. Res.*, Vol. 35, No. 5, pp. 395-403, 2011

[46] J. Zhai, A. B. Barreto, C. Chin, and C. Li, "Realization of stress detection using psychophysiological signals for improvement of human-computer interactions," in *SoutheastCon, 2005. Proceedings. IEEE*, 2005, pp. 415-420

[47] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *Feature extraction, construction and selection*, Springer, 1998, pp. 117-136

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014

[49] M. Z. Hossain and T. Gedeon, "Classifying Posed and Real Smiles from Observers' Peripheral Physiology," in *11th International Conference on Pervasive Computing Technologies for Healthcare*, 2017

[50] L. Chen, T. Gedeon, M. Z. Hossain, and S. Caldwell, "Are you really angry?: detecting emotion veracity as a proposed tool for interaction," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, 2017, pp. 412-416

[51] X. Zhu, Z. Qin, T. Gedeon, R. Jones, M. Z. Hossain, and S. Caldwell, "Detecting the Doubt Effect and Subjective Beliefs Using Neural Networks and Observers' Pupillary Responses," in *International Conference on Neural Information Processing*, 2018, pp. 610-621

[52] X. Li, B. Hu, S. Sun, and H. Cai, "EEG-based mild depressive detection using feature selection methods and classifiers," *Comput. Methods Programs Biomed.*, Vol. 136, pp. 151-161, 2016

[53] J. F. Plested, T. D. Gedeon, X. Y. Zhu, A. Dhall, and R. R. Geocke, "Detection of universal cross-cultural depression indicators from the physiological signals of observers," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* 2017, pp. 185-192

[54] K. A. Buza and J. Koller, "Classification of electroencephalograph data: A hubness-aware approach," *Acta Polytech. Hungarica*, Vol. 13, No. 2, pp. 27-46, 2016

[55] J. Katona and A. Kovari, "Examining the Learning Efficiency by a Brain-Computer Interface System," *Acta Polytech. Hungarica*, Vol. 15, No. 3, pp. 251-280, 2018

[56] M. Tariq, P. M. Trivailo, Y. Shoji, and M. Simic, "Comparison of Event-related Changes in Oscillatory Activity During Different Cognitive Imaginary Movements Within Same Lower-Limb," *Acta Polytech.*

Hungarica, Vol. 16, No. 2, pp. 77-92, 2019

[57] J. Zhai and A. Barreto, "Stress detection in computer users through non-invasive monitoring of physiological signals," *Blood*, Vol. 5, No. 0, 2008

[58] O. Barral, I. Kosunen, and G. Jacucci, "No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment," *ACM Trans. Comput. Interact.*, Vol. 24, No. 6, p. 40, 2018

[59] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Trans. Affect. Comput.*, Vol. 3, No. 3, pp. 298-310, 2012

[60] L. ten Brinke, D. Stimson, and D. R. Carney, "Some evidence for unconscious lie detection," *Psychol. Sci.*, Vol. 25, No. 5, pp. 1098-1105, 2014